Supplementary Methods and Results

1. Supplementary Methods

Culture of biological material and DNA extraction

All 20 *R. irregularis* isolates (Supplementary Table 1) were grown *in vitro* in association with Ri T-DNA transformed carrot roots (Bécard and Fortin, 1988). Roots and isolates were co-cultured in split-plates (St-Arnaud *et al.*, 1996) with a hyphal compartment to allow the collection of fungal material. After 4 months growth at 25°C, hyphal compartments, containing mycelium and spores, were dissolved in citrate buffer (450 ml ddH_2O, 8.5 ml 0.1 N citric acid, 41.5 ml 0.1 N NaCitrate) for 1.5 hours and washed with sterile double-deionized water (ddH_2O). The DNeasy Plant Mini kit (Qiagen) was used for DNA extraction, and DNA was recovered with 60 µl AE buffer. Each DNA sample was extracted from 1 to 4 pooled hyphal compartments. Several independent hyphal compartments of each isolate were used to generate the different biological replicates. The overall DNA extraction procedure was replicated twice (for amplicon sequencing) or three times (for ddRAD-seq) for each isolate. DAOM 197198 was replicated five times for ddRAD-seq.

Saccharomyces cerevisiae, strain S288C, was cultured in the lab of Yves Poirier (Department of Plant Plant Molecular Biology, University of Lausanne, Switzerland) following standard culture conditions, while *Schizosaccharomyces pombe*, strain 972h-, was cultured in the lab of Sophie Martin (Department of Fundamental Microbiology, University of Lausanne, Switzerland) following standard culture conditions. Two independent cultures were used to yield two biological replicates of each yeast species. The genomic DNA of these yeasts was extracted using the DNeasy Plant Mini kit (Qiagen).

Candida albicans, isolates SC5314 and DSY294, was cultured in the lab of Dominique Sanglard (Institute of Microbiology, University Hospital, Lausanne, Switzerland) following standard culture conditions. Three independent cultures of each *C. albicans* strain were established and used to generate 3 replicate DNA samples. The Gentra Puregene DNA extraction kit (Qiagen) was used for DNA extraction.

The genomic DNA of two isolates of *R. irregularis,* B1 and C4, were mixed in known proportions to determine whether the presence of each isolate could be detected after sequencing and whether the frequency of alleles reflected the proportion of DNA mixed. DNA of B1 and C4 were extracted as described above and known amounts of DNA of each isolate were mixed to yield a total DNA concentration of 20 ng/µl. We prepared four mixes of DNA with the following proportions (B1:C4 ratio): 95:5, 70:30, 50:50 and 20:80. These mixes were not replicated.

ddRAD sequencing paired-end library construction and sequencing

All DNA samples were diluted to 20 ng/µl (except the ones that had a lower initial concentration). DNA was digested with *Mse*I and *Eco*RI for 2 h at 37°C, followed by heat inactivation of the enzymes for 20 min at 65°C. Each digestion reaction consisted of 6 µl genomic DNA, 0.9 µl 10 × T4 DNA ligase buffer, 0.45 µl 1 M NaCl, 0.45 µl 1 mg/ml BSA (NEB), 0.1 µl *Mse*I (10,000 U/ml, NEB), 0.25 µl *Eco*RI (20,000 U/ml, NEB) and 0.85 µl ddH_2O.

We prepared barcoded P1-adapters complementary to the *Eco*RI overhang by annealing the following oligonucleotides: EcoRI-P1.1: 5'-CTCTTTCCCTAC-ACGACGCTCTTCCGATCTNNNNNNNC-3' and EcoRI-P1.2: 5'-AATTGNNNN-NNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3', where NNNNNNN denotes the barcode sequence (Supplementary Table 11). We prepared the P2-adapter complementary to the *Mse*I overhang by annealing the following two oligonucleotides: MseI-P2.1: 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCC-GATCT-3' and MseI-P2.2: 5'-Phos-TAAGATCGGAAGAGCGAGAACAA-3'. P2 is a forked adapter, which allows the amplification of only the DNA fragments that ligated to both P1 and P2 adapters. We added 1 µl of a different barcoded-P1 adapter (1 µM) to each digested DNA sample, along with 0.16 µl 10 × T4 DNA ligase buffer, 0.13 µl 1M NaCl, 0.13 µl 1 mg/ml BSA, 1 µl 10 µM P2 adapter, 0.1675 µl T4 DNA ligase (2,000,000 U/ml, NEB) and 0.0125 µl ddH_2O. The adapters were ligated for 6 h at 16°C, followed by enzyme heat-inactivation for 10 min at 65°C. After adapter ligation, we diluted all samples with 100 µl of ddH_2O, and we purified them by using a 1 × ratio of Agencourt AMPure XP magnetic beads (Beckman and Coulter, Inc.) followed by elution with the same volume of ddH_2O.

We amplified adapter-ligated DNA fragments by PCR by using primers complementary to the P1 and the P2 adapters. Each sample was amplified in duplicate with the following conditions: 4 µl purified DNA, 4 µl 5 × Q5 HF DNA polymerase buffer (NEB), 4 µl 5 × high GC enhancer, 0.16 µl 25 mM dNTP, 6.3 µl ddH_2O, 0.2 µl Q5 HF DNA polymerase (2,000 U/ml, NEB) and 1.34 µl of a mix of two PCR primers (5 µM each; Ill-PCR1: 5'- A*A*TGATACGGCGACCACCG-AGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3', Ill-PCR2: 5'-C*A*AGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCT-3'; * denotes phosphorothioate bonds). Cycling conditions were: 30 sec at 98°C, 18 cycles of 20 sec at 98°C, 30 sec at 60°C, 40 sec at 72°C, and a final step for 10 min at 72°C. At the end of the PCR, we added 0.1 µl ddH_2O, 0.4 µl 10 × Q5 HF buffer, 1.34 µl primer mix and 0.16 µl 25 mM dNTP to each PCR reaction, and we ran a final amplification step for 3 min at 98°C, 2 min at 60°C and 12 min at 72°C.

To verify the success of the amplification, we ran 4 µl of each PCR reaction on a 1.5% agarose gel in 1 × TBE, stained with ethidium bromide, for 80 min at 100V. We pooled all PCR reactions into a single tube and purified them with a 1 × ratio of Agencourt AMPure XP magnetic beads followed by elution with 90 µl ddH$_2$O. Library concentration was measured with a Quant-iT PicoGreen dsDNA assay kit (Life Techologies^{TM}) and fragment size distribution was determined on a fragment analyzer (Advanced Analytical) to ensure absence of primer dimer. The library was sequenced on an Illumina HiSeq 2500 device (100 bp paired-end reads) at the Lausanne Genomic Technologies Facility, Switzerland. ddRAD-seq data were deposited in the NCBI SRA database (BioProject Accession Number: PRJNA268659).

ddRAD-seq read processing

We processed the raw reads with the script Tagcleaner.pl to trim Illumina adapters (Schmieder *et al.*, 2010). Reads were quality-filtered and trimmed using PrinSeq-lite.pl version 0.20.4 (Schmieder and Edwards, 2011). Low quality 3'-ends were trimmed and reads containing uncalled bases (N) removed. Only reads longer than 50 bp were kept for further analyses. Reads were then extracted by process_radtags from Stacks v0.9991 (Catchen *et al.*, 2013) according to the barcode sequences and the presence of the *Eco*RI site.

Genome-based prediction and characterization of ddRAD-seq fragments and detection of repeated regions

The technique of RAD sequencing was developed to characterize polymorphisms in populations of non-model species with unknown genome sequences. However, we took advantage of the reference genomes available for all the species that we studied to predict ddRAD-seq fragments *in silico* and to characterize them with respect to their uniqueness in the considered genome. We used two different complementary approaches to identify 'repeated' genomic regions: (1) Recognition of the repeated regions by the de-novo repeat family identification and modeling package RepeatModeler and (2) global pairwise homology among *in silico*-predicted ddRAD-seq fragments to identify paralogs using ggsearch36. The program ggsearch36 was set up to a match/mismatch score of +5/-10 and an E-value threshold of 10^{-20}. Repeated regions were treated separately in our analyses and the conclusions reported in this paper mostly focus on the analysis of reads that map to non-repeated regions in the DAOM 197198 genome. Hence, this represents the most conservative estimate possible of the among- and within-isolate polymorphism in the population of *R. irregularis* isolates without additional genome sequences of other isolates.

Calculation of entropy

To determine the level of genetic polymorphism within isolates, a second independent method, mean Shannon entropy (Shannon, 1948) was calculated for each isolate. Entropy takes into account the number of alleles per site and their frequency.

$$E = -\frac{1}{N_2}\sum_{i=1}^{N_2}\sum_{x} f_{i,x}\cdot \log_2 f_{i,x}$$

where N_2 is the total number of sites per sample, and $f_{i,x}$ is the relative frequency of allele x at position i. For example, a position with a single allele has an entropy of 0, while a position with 4 alleles in equivalent proportions has an entropy of 2. If the 4 alleles are present in different relative proportions, the entropy is lower than 2. An isolate with few poly-allelic sites will have lower entropy than an isolate with many poly-allelic sites.

Haplotype characterization

Within each ddRAD-seq locus, we determined the number of different haplotypes by developing a method to link genetic variants present on individual sequence reads. With custom perl scripts, we first joined aligned paired-end reads into single, joined reads. Second, all variable sites in the population were recorded and only these sites were subsequently used to generate haplotypes. Third, all variants were relocated on joined reads and recorded to generate primary haplotypes. Because sequence reads aligning to a ddRAD-seq locus can have different lengths, some sites on certain haplotypes could not be verified, yielding some missing values (NAs). Filtering and completion of NAs was performed independently at each ddRAD-seq locus: Uninformative haplotypes and sites (i.e. only NAs) were removed and ambiguous haplotypes were resolved by removing the least informative sites; all remaining NAs were imputed by k-Nearest Neighbor Imputation with the R package VIM version 4.0.0 (Templ *et al.*, 2013), using complete haplotypes as references. This step was implemented to ensure no generation of new haplotypes, resulting in a conservative estimate of the number of haplotypes per ddRAD-seq locus and avoiding overestimation. Finally, the calculation of haplotype number was performed at each ddRAD-seq locus for each genome characteristic (coding/non-coding; repeated/non-repeated). Only ddRAD-seq loci with a depth of coverage greater or equal to 10 were considered. The number of haplotypes at each ddRAD-seq locus was counted and the proportion of ddRAD-seq loci having a given number of haplotypes was established. ddRAD-seq loci including at least one repeated region position or one coding region position were assigned to the “repeated” category or to the “coding” characteristic category, respectively.

Calculation of synonymous and non-synonymous mutations

The ratio of substitution rates at non-synonymous and synonymous sites (K_a/K_s) is difficult to interpret within a population (Kryazhimskiy and Plotkin, 2008) and this type of analysis is problematic with ddRAD-seq data. Using custom perl scripts, we defined partial transcript sequences based on overlaps of exon sequences detected by GeneMark-ES, of predicted ddRAD-seq fragments and of haplotypes characterized from the ddRAD-seq data. We verified the translation frame of each partial transcript, and conserved only partial transcripts that contained no indels (because they cause frameshift mutations), that were non-repeated and that were present in all isolates. Only sites with 2 different codons were considered for analysis. Sites with 3 or more different codons and sites with a unique codon were rejected. A codon pair was defined as synonymous if the 2 codons were synonymous, and defined as non-synonymous if the 2 codons were non-synonymous, based on the standard genetic code. The differences in proportions of non-synonymous codon pairs between isolates or strains were tested with a Chi-squared test ($p < 0.001$). This analysis was performed on all *R. irregularis* isolates, *Betula* spp. and *C. albicans* isolates.

Amplicon sequencing

We amplified all loci (Supplementary Table 4) using the following PCR conditions: 2 ng of genomic DNA, 2.5 µl 5 × Q5 HF DNA polymerase buffer, 0.1 µl 25 mM dNTP, 2.5 µl 5 × high GC enhancer, 2.5 µl 1 µM primer "For", 2.5 µl 1 µM primer "Rev", 0.2 µl Q5 HF DNA polymerase (2,000 U/ml, NEB) and 1.2 µl ddH2O. Cycling conditions were: 4 min at 98°C, 30 cycles of 30 sec at 98°C, 1 min at 60°C, 1 min at 72°C, and a final step for 2 min at 72°C. Amplification success was verified by running 3 µl of each PCR reaction on a 1% agarose gel in 1 × TBE, stained with ethidium bromide, for 1 h at 100 V. PCR products were pooled into one tube. We concentrated the pooled DNA by ethanol precipitation and purified it with a 1.5 × ratio of Agencourt AMPure XP magnetic beads followed by elution with 50 µl ddH_2O. The DNA was then converted into a sequencing-compatible library following the TruSeq DNA library preparation protocol (Illumina) without DNA fragmentation. This was then repeated using independent DNA samples of the same isolates. Both libraries were sequenced on separate lanes of Illumina MiSeq sequencer (250 bp paired-end reads) at the Lausanne Genomic Technologies Facility.

2. Supplementary Results

Characterization of the different genomes

We predicted the ddRAD-seq fragments in each genome based on the presence of the restriction sites and on the length of the fragments. This strategy allowed us to work on known locations in the genomes and to characterize precisely the predicted ddRAD-seq fragments. *In silico* digestion and size filtering generated 6427, 36235 and 88374 predicted ddRAD-seq fragments in the genome of *C. albicans*, *R. irregularis* and *B. nana*, respectively (Supplementary Table 2). Among these fragments, 1%, 1% and 8% of the fragments could not be unambiguously relocated in the genome and were rejected. The clustering strategy using ggsearch36 revealed that

3%, 12% and 13% of the predicted ddRAD-seq fragments are not single copy elements in the genome of *C. albicans*, *R. irregularis* and *B. nana*, respectively.

Approximately 62%, 31% and 15% of the genomes of *C. albicans*, *R. irregularis* and *B. nana*, were found to be coding (Supplementary Table 2). Approximately 4%, 30% and 34% of the genomes of *C. albicans*, *R. irregularis* and *B. nana*, were found to contain repeated elements identified by RepeatModeler/RepeatMasker (Supplementary Table 2).

Analysis of ddRAD-seq data

After quality filtering with Tagcleaner.pl and PrinSeq-lite.pl, a total of 350 778 768 reads were de-multiplexed, resulting in 108 896 313 high quality reads that were retained for further analyses (Supplementary Table 5). The mean number of predicted ddRAD-seq fragments with coverage ≥ 10× was 17 232 (±301 SE, range 10 841-21 415). Results for other species are also reported in Supplementary Table 5.

Calculation of consistency among replicates

Consistency of a SNP existed at a site when all biological replicates of a given isolate showed exactly the same SNP composition. For example, 3 replicates sharing an adenine (A / A / A pattern) at a given site had a consistent SNP site. If the pattern was A / A / G, the site was recorded as inconsistent. We distinguished the absence of data from inconsistency: a missing allele, for example because of low coverage, in one of the replicates (A / A / na pattern) was not considered as inconsistent. Consistency of mono-allelic SNPs was at least 99% in all isolates from the Swiss population (Supplementary Table 7), showing the very good quality of the data. For poly-allelic SNPs, the consistency was calculated as follows. At one site, if 3 replicates showed T,A / T,A / T,A, the site was recorded as consistent. We also determined the number of sites with a missing allele in one of the replicates, such as T,A / T,A / T, and took these sites into account to calculate the inconsistency level. In this situation, the consistency was dependent on the missing allele, since we could not exclude that the missing "A" allele was absent because of technical issues (e.g. low coverage) or because it did not exist in the sample. Finally, if 3 replicates showed T,A / T,A / T,C, this site was recorded as inconsistent. The percentages of consistency ranged from 70.8% to 93.1% for the poly-allelic sites in the Swiss population. Even if these percentages were lower than the consistency of mono-allelic SNPs, they remained high and were unlikely explained by artifacts.

Allele frequency distribution in DNA mixes

We mixed the DNA of isolates B1 and C4 in known proportions to determine whether allele frequencies reflected the proportions of each DNA. We determined allele frequencies by counting the number of reads covering each allele at sites that were mono-allelic in both B1 and C4, but could become bi-allelic in the DNA mixes. When all positions were included (*i.e.* repeated/non-repeated and coding/non-coding), allele frequency distributions generally reflected well the proportions of each DNA (Supplementary Figure 12). Most sites remained mono-allelic in the 95:5 mix, with a small proportion of sites with alleles at low frequency. Allele frequencies of the 70:30 and 20:80 mixes both peaked around 0.25 and 0.75. Finally, the allele frequency of the 50:50 mix nicely peaked at 0.5. Allele frequency distributions of non-repeated sites were similar (Supplementary Figure 12).

Poly-allelic SNP density with a 30× coverage threshold

To ensure that poly-allelic SNP density was not over-estimated by the apparently low coverage threshold we chose (10×), we calculated poly-allelic SNP density including only sites covered at least 30× (close to the overall average coverage per locus of 37×, Supplementary Table 5) and a minimum allele frequency threshold of 0.1. Poly-allelic SNP density per isolate was slightly higher when a coverage threshold of 30× was used than with a threshold of 10× (Supplementary Figure 13). When an even higher coverage threshold was used (e.g. 50×), poly-allelic SNP density per isolate was even higher than with 30× as a threshold (data not shown). Therefore, with a high coverage threshold, poly-allelic SNP density was over-estimated because a vast number of mono-allelic sites at low coverage were left out of the analysis. This is exemplified by one of the biological replicates of DAOM 197198: it had the lowest average coverage per locus out of the five replicates, but the highest poly-allelic SNP density out of the five replicates. Choosing a coverage threshold of 10× was a conservative approach to analyze this particular ddRAD-seq data. A higher threshold of 50× was not appropriate given the overall average coverage per locus of 37×; too many interesting loci would have been left out of the analyses.

Mean entropy of the genetic diversity at each site

We used entropy as an independent measure of genetic diversity within an isolate. Mean entropy was 0.033, 0.064, 0.109 and 0.148 bits in non-repeated – non-coding, non-repeated – coding, repeated – non-coding and repeated – coding regions, respectively (Supplementary Figures S8A-E). DAOM 197198 displayed the lowest mean entropy (0.033 bits) and Swiss *R. irregularis* isolates ranged from 0.048-0.115 bits (Supplementary Figure S8B). Mean entropy gave comparable results to the measure of poly-allelic SNP densities.

Minimum haplotype analysis

In non-repeated – coding regions, 97.6% of sequenced ddRAD-seq loci in DAOM 197198 were of only 1 haplotype (Figure 6A, Supplementary Table 9). Among the Swiss isolates, A1 was the isolate with the highest proportion of ddRAD-seq loci with 1 haplotype (95.5%), while C3 was the one with the lowest proportion of ddRAD-seq loci with 1 haplotype (80.4%). C3 had 14.9%, 2.7%, 1.1% and 0.9% of ddRAD-seq loci with 2, 3, 4 or ≥5 haplotypes, respectively (Figure 6A). The diploid *Betula* spp.

revealed that 22% of ddRAD-seq loci had 2 haplotypes and 2-3% had more than 2 haplotypes. Tri- and tetraploid *Betula* spp. comprised 37% of ddRAD-seq loci where there were 2 haplotypes, 15% with 3 haplotypes, 4% with 4 haplotypes, and between 1 to 3% with more than 4 haplotypes (Figure 6B, Supplementary Table 9). *C. albicans* exhibited between 43% and 56% of ddRAD-seq loci with 2 haplotypes, and 6% with more than 3 haplotypes (Figure 6C, Supplementary Table 9). In the haploid yeasts *S. cerevisiae* and *S. pombe*, proportions of loci with 1 haplotype were 100%. The number of haplotypes per ddRAD-seq locus was also calculated in repeated regions, either coding or non-coding, and in non-repeated, non-coding regions (Supplementary Figure 9A-9C). The total number of ddRAD-seq loci analyzed in each sample depends on the genetic variability present within each population. Therefore, proportions of sites with 1, 2 or more haplotypes can be compared only among individuals within a species, but not among species, i.e., haplotype absolute proportions in *R. irregularis* isolates cannot be compared to the ones found in *C. albicans* or *Betula* spp. However, the presence of 1, 2 or more haplotypes depends on the ploidy level or number of genetically different nuclei within each organism, and this result in the *R. irregularis* population would not be expected if the fungus were homokaryotic. In the diploid strains, *C. albicans* and *B. nana*, we detected loci with more than 2 haplotypes at low frequencies (Figure 6B-6C). These may be caused by repeats left undetected by filtering, by copy number variants or by tandem duplications.

Proportions of synonymous and non-synonymous substitutions in poly-allelic sites

In an attempt to evaluate if the within-isolate polymorphism could have functional consequences, we considered partial transcript sequences from coding non-repeated regions and calculated the proportion of non-synonymous codons when the partial transcript revealed 2 haplotypes. In sites with 2-allelic codons, most of the codons that were different from each other were synonymous codons in all samples and species (ranging from 83% to 99%, Figure 7 and Supplementary Figure 10). In *Betula* spp., the proportions of non-synonymous codon pairs were 5% and 6% for the diploid 097-10 and 582 respectively (Figure 7). These 2 proportions were significantly different (Chi-squared test, p < 0.001) from the proportions of non-synonymous codon pairs observed in tri- or tetra-ploids (ranging from 15% to 17%). Proportions of non-synonymous codon pairs observed in the diploid *C. albicans* strains were significantly greater than in the diploid *B. nana* (Chi-squared test, p < 0.001) and slightly lower than in the triploid and tetraploid *Betula* spp. Isolates DAOM 197198, A1, A2 and C2 had significantly lower proportions (ranging from 1% to 4%) of non-synonymous codon pairs than the diploid *B. nana* 097-10 (Chi-squared test, p < 0.001), C3 had 9% of non-synonymous codon pairs, which was significantly greater than in DAOM 197198, A1, A2, C2 and the diploid *B. nana*, but not significantly different from *C. albicans* strains (Figure S10). C3, A4 and C4 isolates showed the greatest proportions (ranging from 6% to 12%) of non-synonymous codon pairs when compared to all other isolates (Supplementary Figure 10).

References

Catchen J, Hohenlohe P a, Bassham S, Amores A, Cresko W a. (2013). Stacks: an analysis tool set for population genomics. *Mol Ecol* **22**:3124–40.

Kryazhimskiy S, Plotkin JB. (2008). The population genetics of dN/dS. *PLoS Genet* **4**:e1000304.

Lin K, Limpens E, Zhang Z, Ivanov S, Saunders DGO, Mu D, *et al.* (2014). Single Nucleus Genome Sequencing Reveals High Similarity among Nuclei of an Endomycorrhizal Fungus. *PLoS Genet* **10**:e1004078.

Schmieder R, Edwards R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**:863–4.

Schmieder R, Lim YW, Rohwer F, Edwards R. (2010). TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* **11**:341.

Shannon CE. (1948). A mathematical theory of communication. *Bell Syst Tech J* **27**:379–423,623–656.

Templ M, Alfons A, Kowarik A, Prantner B. (2013). VIM: Visualization and Imputation of Missing Values. R package version 4.0.0. http://cran.r-project.org/package=VIM.